Classification de codecs de la parole et du son sur des critères perceptuels

YVES ZANGO[1,2,3], RÉGINE LE BOUQUIN JEANNÈS[2,3], NATHALIE COSTET[2,3], CATHERINE QUINQUIS[1]

[1]Orange Labs R&D TECH/OPERA - 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2]INSERM, U 642, Rennes, F-35000, France
[3] Université de Rennes 1, LTSI, F-35000, France
LTSI, Campus de Beaulieu, Université de Rennes 1, 35042 Rennes Cedex, France

[1,2,3] yves.zango@orange-ftgroup.com, catherine.quinquis@orange-ftgroup.com,
[2,3] regine.le-bouquin-jeannes@univ-rennes1.fr, [2,3] nathalie.costet@univ-rennes1.fr

Résumé - Dans le cadre du développement d'un système de signaux de référence pour l'évaluation subjective de la qualité audio et vocale des codeurs actuels, nous proposons une classification hiérarchique, basée sur des critères perceptuels, des codeurs nouvelle génération dont la bande de fréquences est limitée à [50 Hz - 7000 Hz]. Pour ce faire, nous avons réalisé un test de dissimilarité sur un ensemble de codeurs large bande. Les matrices de dissimilarités obtenues à l'issue du test ont été traitées par deux techniques de réduction de dimensions. A partir des espaces perceptifs générés par ces deux techniques, une classification hiérarchique a permis d'une part de mettre en évidence une forte corrélation entre les techniques de codage des codecs et leur qualité perceptive et d'autre part de comparer les performances des deux approches.

Abstract - In the objective of the design of reference signals for current audio and voice codecs subjective quality assessment, we propose a hierarchical clustering based on perceptual criteria of new generation codecs whose bandwidth is limited to the frequency band [50 Hz - 7000 Hz]. To do this, we performed a dissimilarity test on a set of wideband codecs. The dissimilarity matrices produced by this test were analyzed using two dimensionality reduction techniques. From the perceptual spaces derived from both analyses, a hierarchical classification allowed us, on the one hand, to highlight a strong correlation between coding techniques and codecs perceptual quality and, on the other hand, to compare the results from these two techniques.

1 Contexte

L'essor des télécommunications se traduit aujourd'hui par la naissance de nouveaux services tels que la VoIP (Voice over IP), la vidéoconférence et, conjointement, au cours des dernières années, de nouveaux codecs audio ont vu le jour. Au vu de la haute concurrence qui règne dans le monde des télécommunications, la qualité du service proposé devient un enjeu majeur. Afin de proposer à leurs clients une qualité audio satisfaisante des différents services vocaux, les laboratoires de recherche de nombreux opérateurs de télécommunications doivent procéder à des évaluations objective et/ou subjective de leurs équipements. Ils évaluent notamment la qualité des codecs qu'ils utilisent. L'évaluation objective des codecs modélise une note de qualité, tandis que l'évaluation subjective vise à attribuer une note aux codecs en se basant sur le jugement d'un groupe d'auditeurs entraînés ou non. Notre étude porte sur l'évaluation subjective de la qualité intrinsèque des nouveaux codecs dits à bande élargie ([50 Hz - 7000 Hz]), limitée à la parole claire, les autres signaux tels que la parole bruitée ou la musique n'étant pas pris en compte. De plus, les problèmes liés à la transmission, erreurs binaires ou trames effacées, ne seront pas traités dans cette étape de l'étude. Dans un premier temps, il s'agit de sélectionner un ensemble de codecs large bande présentant différentes techniques de compression, et de faire passer à des auditeurs un test de dissimilarité visant à comparer par paires ces codecs. Aux matrices de dissimilarité obtenues, nous proposons d'appliquer deux types d'analyse de réduction de dimension afin de ne retenir que les dimensions principales caractérisant l'espace perceptif des codecs et d'effectuer une classification hiérarchique sur les résultats obtenus par ces deux techniques, pour comparer leurs performances. L'objectif visé est d'examiner la cohérence des deux techniques, l'idée étant de retenir par la suite la plus pertinente dans la phase future d'élaboration de signaux d'ancrage.

2 Codecs et techniques de codage

L'étude est basée sur des codecs nouvelle génération ayant une largeur de bande supérieure à la traditionnelle bande étroite ([300 Hz – 3400 Hz]). Afin de couvrir les diverses techniques utilisées actuellement en compression de la parole, nous avons sélectionné des codecs utilisant différentes techniques de codage résumées dans le Tableau 1. Nous avons distingué cinq groupes de techniques de codage. Le premier groupe est composé des codecs G.722.1 [1] et G.722.1C [2] (l'annexe C du codec G.722.1) qui utilisent la technique de codage par transformation à modulation et chevauchement ou MLT (Modulated Lapped Transform). Le second est celui des codecs G.722.2 [3] qui utilisent la technique de prédiction linéaire avec excitation par code algébrique ou ACELP (Algebraic Code-Excited Linear Predictive). Le troisième groupe correspond à un codage par forme d'onde (codecs G.722 [4]). Le quatrième groupe comprend des codeurs hybrides (codecs G.729.1 [5]) qui utilisent la technique de codage par transformée en cosinus discret modifié ou MDCT (Modulated Discrete Cosine Transform) et la technique ACELP. Le dernier groupe est celui des codeurs MDCT (codecs MP3 [6] et HEAAC [7]). Les codecs des premier et dernier groupes sont qualifiés de codecs par transformées. Tous ces codecs fonctionnent à différents débits, et il a été convenu de retenir dans cette étude 19 codecs présentant différents débits. Nous avons appliqué un transcodage d'ordres 2 et 3 (tandems) sur ces codecs, et un test ACR (Absolute Category Rating) [8] a été réalisé sur les 58 conditions ainsi obtenues (les 3 tandems pour chaque codec plus la condition initiale correspondant au signal original). A l'issue du test ACR, seuls 20 codecs/tandems ont été conservés. Ces codecs obtenaient une note MOS (Mean

Opinion Score) entre 2,5 et 3,5 sur une échelle de 5, ceci afin que les jugements de dissimilarité portent plus sur la perception des dégradations que sur la qualité globale. Les 20 codecs/tandems sélectionnés sont présentés dans le Tableau 2. Les symboles x2 et x3 signifient que la fonction de transcodage est appliquée respectivement 2 et 3 fois. Le symbole x1 signifie que le codeur n'a subi aucun transcodage.

Tab. 1 : Techniques de codage

Codecs	Caractéristiques techniques
G722.1	Codage MLT (Modulated Lapped Transform)
G.722.1C	
G.722.2	Codage ACELP (Algebraic Code Excited Linear Prediction)
G.722	Codage par forme d'onde (Waveform coding)
G.729.1	Codage hybride (Hybrid coding)
HEAAC	Codage MDCT (Modulated Discrete Cosine Transform)
MP3	

Tab. 2 : Liste des codecs/tandems retenus

Indice	Description	Indice	Description
1	G722.1C_24kbps_x2	11	G722_56kbps_x2
2	G722.1C_24kbps_x3	12	G722_56kbps_x3
3	G722.1_24kbps_x2	13	G729.1_14kbps_x3
4	G722.1_24kbps_x3	14	G729.1_20kbps_x3
5	G722.2_12.65kbps_x2	15	G729.1_24kbps_x2
6	G722.2_12.65kbps_x3	16	G729.1_32kbps_x3
7	G722.2_15.85kbps_x2	17	HEAAC_24kbps_x2
8	G722.2_8.85kbps_x2	18	HEAAC_32kbps_x2
9	G722_48kbps_x2	19	MP3_32kbps_x1
10	G722_48kbps_x3	20	MP3_32kbps_x2

3 Procédure expérimentale

Les signaux originaux utilisés pour l'expérimentation sont des doubles-phrases prononcées par un homme et par une femme. Ces phrases ont une durée totale de 6 secondes et sont séparées par un court silence. Les stimuli utilisés pour l'étude correspondent à ces doubles-phrases traitées par les 20 codecs/tandems. Certains codecs utilisés étant des codecs à bande super élargie, une limitation de bande à la largeur de bande [50 – 7000 Hz] a été appliquée aux stimuli, afin d'éviter une influence de cette largeur de bande sur la qualité perceptive. Un test de dissimilarité a été construit à partir de ces stimuli, durant lequel on a demandé à 29 auditeurs de noter sur une échelle continue de 0 (l'auditeur juge que les stimuli sont absolument identiques) à 100 (les stimuli sont jugés complètement différents), la distance qu'ils percevaient entre les 210 paires de stimuli proposées à l'écoute (C_{20}^2+10 paires nulles).

4 Analyse statistique

Une fois les matrices de dissimilarité obtenues, nous leur avons appliqué deux techniques de réduction de dimension, permettant de mettre en évidence les principales dimensions décrivant l'espace perceptif des codecs. Les dimensions de ces espaces représentent les principaux défauts perceptifs introduits par les techniques de codage implémentées dans les codecs considérés. Ces deux techniques d'analyse sont, d'une part, une analyse multidimensionnelle INDSCAL (INdividual Differences SCALing) et, d'autre part, une Analyse Factorielle Multiple (AFM).

4.1 Modèle INDSCAL

Le modèle INDSCAL est une analyse multidimensionnelle (MDS : MultiDimensional Scaling) [9] appliquée à plusieurs matrices de dissimilarité. Elle permet de projeter les différentes matrices de dissimilarité dans un espace euclidien commun. La distance euclidienne entre les points de cet espace vise à approximer les dissimilarités initiales des stimuli. Supposons m matrices de dissimilarité de taille $n \times n$ correspondant aux matrices obtenues à partir de m sources (qui correspondent ici aux auditeurs) comparant n objets (en l'occurrence les stimuli) par paire. Soit S la dimension de l'espace commun. En considérant uniquement la $k^{ème}$ source, le modèle INDSCAL peut être reformulé en écrivant la distance euclidienne entre les objets i et j comme suit :

$$d_k^2(i,j) = \sum_{s=1}^{S} q_s^k \left(z_s(i) - z_s(j) \right)^2 + e_k(i,j) .$$

Dans l'expression ci-dessus $e_k(i,j)$, $z_s(i)$ et q_s^k correspondent respectivement à l'erreur d'approximation, la coordonnée de l'objet i sur la $s^{ème}$ dimension et le poids accordé par la $k^{ème}$ source à la $s^{ème}$ dimension. Comme d_k est une distance euclidienne, le produit scalaire associé peut être défini par :

$$\langle i | j \rangle_k = \sum_{s=1}^{S} q_s^k z_s(i) z_s(j) + \varepsilon_k(i,j),$$

d'où découle la matrice du produit scalaire :

$$M_k = Z Q_k Z^T + E_k .$$

Le problème de minimisation de l'erreur d'approximation E_k est résolu via différents algorithmes tels que ALSCAL (Alternating Least squares SCALING) ou PROXSCAL (PROXimity SCALing), ce dernier étant l'algorithme retenu dans cette étude.

4.2 PROXSCAL

Soient $\Delta_k = (\delta_{ij}) \in \mathbb{R}^{n \times n}$ la matrice de dissimilarité de la $k^{ème}$ source. L'algorithme PROXSCAL cherche itérativement la configuration $X_k \in \mathbb{R}^{n \times p}$ de sorte que la distance $d(X_k)$ entre les n objets approche au mieux la dissimilarité Δ_k. La variable $d(X_k)$ correspond à la distance entre les lignes de X_k, généralement euclidienne. La fonction d'erreur, appelée *stress*, que l'algorithme vise à minimiser s'écrit :

$$\sigma = \frac{1}{m} \sum_{k=1}^{m} \sum_{i<j}^{n} w_{ijk} \left[\hat{\delta}_{ij} - d_{ij}(X_k) \right]^2$$

où $\hat{\delta}_{ij} = f(\delta_{ij})$, appelée disparité, est une régression monotone de la dissimilarité δ_{ij} et w_{ijk} est le poids accordé par la source k à la dissimilarité δ_{ij}. Selon la nature de la fonction f, nous pouvons distinguer plusieurs formes de l'algorithme. Si f est une fonction linéaire, on qualifie la MDS de « métrique ». Sinon, pour toute autre fonction f monotone, on la qualifie de « non métrique ». Dans la mesure où les rangs des dissimilarités sont jugés plus fiables que les

valeurs elles-mêmes [10], nous avons opté pour la MDS non métrique. Ainsi, ce sont les rangs des dissimilarités qui ont été considérés et non les dissimilarités elles-mêmes. Les coordonnées des objets sont calculées pour des configurations allant de 2 à p dimensions (p étant fixé a priori par l'utilisateur). Pour chaque configuration, le stress est calculé et l'algorithme s'arrête lorsque l'amélioration du stress est inférieure à un seuil fixé. Le choix du nombre optimal de dimensions $\left(p^*\right)$ est obtenu en analysant la courbe du stress final obtenu pour chaque configuration.

4.3 AFM

A partir de n individus décrits par m groupes de variables $V_j, j \in \{1,\cdots,m\}$, on construit m tableaux $X_j, j \in \{1,\cdots,m\}$ du type $individus \times variables$. Dans notre étude, les n individus sont les codecs et les m tableaux représentent les espaces perceptuels des m auditeurs. L'AFM vise à trouver un espace de projection commun aux auditeurs en prenant en compte à la fois tous les groupes de variables. Pour ce faire, une première Analyse en Composantes Principales (ACP) est appliquée à chacun des tableaux, puis les coordonnées des codecs dans l'espace commun sont obtenues en appliquant une seconde ACP (ACP globale) sur l'ensemble des tableaux pondérés par la racine carrée de la plus grande valeur propre obtenue à l'issue de la première ACP. Le nombre de dimensions optimal se déduit de la courbe des valeurs propres issues de l'analyse globale. Dans notre cas, afin de pouvoir appliquer l'AFM, les matrices de dissimilarités sont préalablement transformées en matrices de distance euclidienne avant d'être transformées en matrices (ou tableaux) $individus \times variables$ via une ACP préliminaire. Cette technique est appelée AFMTD (Analyse Factorielle Multiple sur Tableau de Distances) [11]. Dans notre étude, nous avons appliqué une régression monotone aux matrices de dissimilarités avant d'appliquer l'AFM. Cette régression monotone a pour but de transformer les dissimilarités qui sont des données « métriques » en des données « non métriques », et ce pour ne tenir compte que de l'ordre des dissimilarités comme dans le cas de l'approche précédente (par PROXSCAL).

4.4 AFM et modèle INDSCAL

Un avantage de l'AFM est qu'elle est exempte de tout problème de convergence (dans la mesure où il s'agit d'une double ACP qui est une technique basée sur une diagonalisation de matrice) contrairement aux autres algorithmes (PROXSCAL et ALSCAL) qui peuvent parfois être confrontés à l'existence de minima locaux (lors des itérations visant à minimiser le stress). D'autre part, l'AFM permet d'intégrer des variables supplémentaires dans l'analyse. Ces variables n'interviennent pas dans la construction de l'espace commun, mais elles permettent de décrire cet espace. Par ailleurs, l'AFM permet de traiter simultanément des tableaux de nature et de statut différents. Ainsi, il est possible d'inclure dans l'analyse des groupes de variables qualitatives (par exemple, des éléments de verbalisation) et de leur donner un statut d'éléments supplémentaires illustratifs permettant d'interpréter a posteriori le sens de la structure mise en évidence sans qu'ils ne soient intervenus comme éléments constructeurs de cette structure.

5 Classification ascendante hiérarchique

A partir des coordonnées des codecs obtenues dans les analyses visant à réduire le nombre de dimensions (INDSCAL et AFM), nous avons appliqué une classification ascendante hiérarchique (CAH) basée sur le critère d'agrégation de Ward. Cet algorithme vise à minimiser la distance de Ward définie comme suit :

$$D_{k,k'} = \frac{m_k m_{k'}}{m_k + m_{k'}} \left\| x_{G_k} - x_{G_{k'}} \right\|^2_{\mathbb{R}^p} .$$

Dans l'expression ci-dessus, k et k' correspondent aux indices des classes à agréger, G_k et $G_{k'}$ leurs centroïdes respectifs, m_k et $m_{k'}$ leurs masses respectives, x_{G_k} et $x_{G_{k'}}$ leurs coordonnées respectives dans l'espace à p dimensions. Cette classification permet de construire une hiérarchie de partitions et d'observer sous forme d'un arbre hiérarchique le processus d'agrégation des codecs. Le choix du nombre pertinent de classes à retenir repose sur l'observation de cet arbre et des indices d'agrégation correspondants. L'interprétation des classes se fait en observant les codecs qui les composent, ainsi que d'éventuels éléments supplémentaires issus de l'AFM.

6 Résultats

6.1 Choix du nombre optimal de dimensions

Le choix du nombre optimal de dimensions se fait en analysant la courbe du stress (appelée aussi « scree plot ») dans le cas d'INDSCAL et la courbe des valeurs propres dans le cas de l'AFM. Dans le cas d'INDSCAL, la Figure 1 montre un léger « coude » entre les 4ème et 5ème dimensions, ce qui nous incite à choisir 4 comme nombre optimal de dimensions pour cette analyse.
De même, la courbe des valeurs propres issues de l'analyse AFM (Figure 2) révèle un « coude » assez net au niveau de la 5ème dimension. Par conséquent, l'analyse AFM permet elle aussi de conclure à un espace perceptif à 4 dimensions.

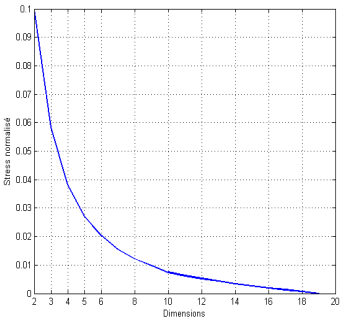

Figure 1 : Courbe du stress – Analyse INDSCAL

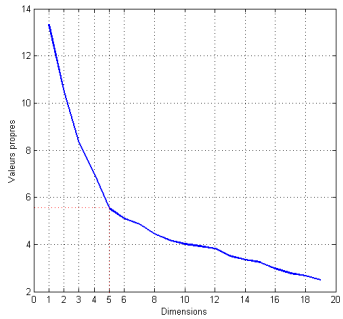

Figure 2 : Courbe des valeurs propres – Analyse AFM

6.2 Description des classes des codecs

Les dimensions des espaces fournis par les deux techniques sont fortement corrélées. Les dendrogrammes des Figures 3 et 4 représentent une CAH appliquée aux coordonnées des stimuli respectivement dans les espaces fournis par INDSCAL et par l'AFM. On constate que, quelle que soit la méthode statistique, 5 groupes de stimuli se forment à l'issue de la CAH. En revanche, selon la méthode statistique choisie, la classification reflète ou non les principales techniques de codage. En effet, les résultats donnés par la méthode INDSCAL mettent en évidence les groupes suivants : les codecs en forme d'onde (codecs 9, 10, 11 et 12), les codecs MLT (codecs 1, 2, 3 et 4), les codecs MDCT (codecs 17, 18, 19 et 20), les codecs ACELP (5, 6, 7, 8) et enfin les codecs hybrides (codecs 13, 14, 15 et 16). Pour ce qui est de l'analyse par AFM, elle conduit à un regroupement en 5 classes différentes de celles précédemment trouvées par INDSCAL. Le premier groupe est composé des codecs en forme d'onde (codecs 9, 10, 11 et 12). Le second groupe comprend des codecs par transformées (codecs 1, 2, 3, 4) auxquels se joignent les codecs hybrides de meilleure qualité (codecs 15 et 16). Cela peut s'expliquer par le fait que les codeurs hybrides possèdent une couche de codage par transformées. Les autres codecs par transformées (codecs MDCT 17, 18, 19 et 20) forment le troisième groupe. Le quatrième groupe est composé des codeurs ACELP (codecs 5, 6, 7 et 8) et finalement le dernier groupe est composé des codecs hybrides de qualité inférieure (codecs 13 et 14).

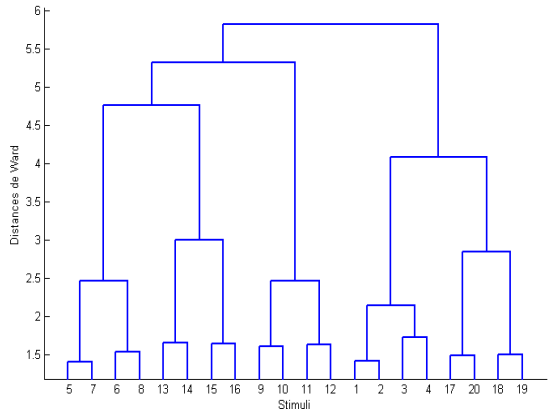

Figure 3 : Dendrogrammes basés sur l'analyse par INDSCAL

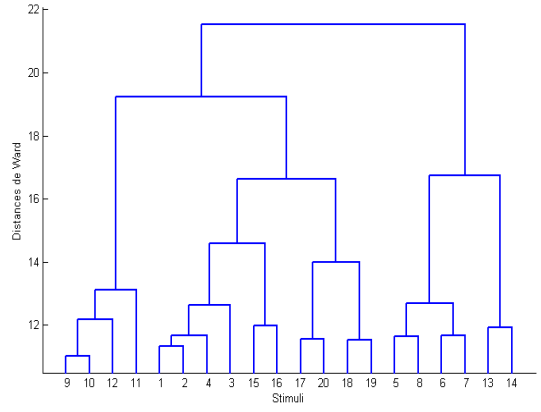

Figure 4 : Dendrogrammes basés sur l'analyse par AFM

7 Conclusions et perspectives

Si, selon la méthode statistique utilisée, la classification hiérarchique ne permet pas de retrouver un regroupement « fidèle » des codecs par technique de codage, les résultats obtenus peuvent se justifier assez facilement. Le fait que les codecs hybrides se regroupent avec les codecs par transformées lorsqu'ils sont utilisés à haut débit n'est pas surprenant dans la mesure où ces codecs comprennent une dernière couche par transformée. Cette couche par transformée a pour but de gommer les défauts persistants du codage CELP en bande étroite et d'améliorer le codage de la bande haute. Ce faisant, les défauts du codage par transformée prennent le dessus. Ainsi, en utilisant l'AFM, les codecs par MDCT se séparent en deux : les mieux notés (en termes de MOS) entrent dans la classe des transformées, les autres formant un groupe à part.

Un test de verbalisation a été mené afin de labelliser les dimensions de l'espace perceptif. Grâce à la possibilité d'analyse simultanée des variables qualitatives et quantitatives que fournit l'AFM, nous pourrons dans une future étude projeter les attributs retenus lors du test de verbalisation dans l'espace perceptif dans le but de qualifier ces dimensions. Ces dimensions seront ensuite modélisées par des signaux qui seront utilisés comme signaux d'ancrage durant les tests d'évaluation de la qualité subjective des codeurs de la parole et du son. Ce nouveau système de signaux d'ancrage devrait remplacer le système MNRU (Modulated Noise Reference Unit) normalisé par l'UIT (Union Internationale des Télécommunications) [12] et actuellement utilisé lors des tests d'évaluation subjective de la parole.

Références

[1] ITU-T Recommendation G.722.1, "Low-complexity coding at 24 and 32 Kbit/s for hands-free operation in systems with low frame loss," 2005.

[2] C. Lamblin, C. Quinquis and P. Usai, "ITU-T G.722.1 annex C: the first ITU-T superwideband audio coder," *IEEE Communication Magazine*, vol. 46, pp.116-122, 2008.

[3] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 Kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," 2003.

[4] ITU-T Recommendation G.722, "7 kHz audio-coding within 64 Kbit/s," 1988.

[5] ITU-T Recommendation G.729.1, "G.729-based embedded variable bit-rate coder: An 8-32 Kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.

[6] T. Sakamoto, M. Taruki, T. Hase, "A fast MPEG-audio layer III algorithm for a 32-bit MCU," *IEEE Transactions on Consumer Electronics* vo1. 45, pp.986-993, no. 3, August, 1999.

[7] J. Herre and J. M. Dietz, "MPEG-4 high-efficiency AAC coding," *IEEE Signal Processing Magazine*, *vol. 25*, no. 3, pp. 137-142, May 2008.

[8] ITU-T Recommendation P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[9] I. Borg, P J.F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, 2nd Edition, Springer, New York, 2005.

[10] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *Acoustical Society of America*, vol. 110, pp. 2167-2182, 2001.

[11] B. Escofier, J. Pagès, *Analyses factorielles simples et multiples, Objectifs, méthodes et interprétation*, 4ème édition, Septembre 2008.

[12] ITU-T Recommendation P.810 "Modulated Noise Reference Unit (MNRU)," February, 1996.